DOCUMENT RESUME

ED 466 639 TM 034 249

AUTHOR Bishop, N. Scott; Omar, Md Hafidz

TITLE Comparing Vertical Scales Derived from Dichotomous and Polytomous IRT Models for a Test Composed of Testlets.

PUB DATE 2002-04-00

NOTE 43p.; Paper presented at the Annual Meeting of the National Council on Educational Measurement (New Orleans, LA, April 2-4, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS Achievement Tests; Comparative Analysis; Elementary Education; *Elementary School Students; *Item Response Theory; *Test Construction; Test Items

IDENTIFIERS Dichotomous Variables; Iowa Tests of Basic Skills; Partial Credit Model; Polytomous Variables; Rasch Model; *Testlets; Three Parameter Model; Vertical Sorting

ABSTRACT

Previous research has shown that testlet structures often violate important assumptions of dichotomous item response theory (D-IRT) models, applied to item-level scores, that can in turn affect the results of many measurement applications. In this situation, polytomous IRT (P-IRT) models, applied to testlet-level scores, have been used as an alternative. The objective of this study was to examine the distributional characteristics of vertical scales created using selected D-IRT and P-IRT models for a test composed of testlets. Iowa Tests of Basic Skills (ITBS) Reading Comprehension test scores from 60,000 randomly selected students (10,000 students per grade for grades 3 through 8) who took the ITBS during a recent fall administration were used in this study. For this data, vertical scales were produced using three D-IRT models (the Rasch model and the three-parameter logistic model calibrated under concurrent and separate group options) and four P-IRT models (the nominal model, the graded response model, the partial credit model, and the generalized partial credit model). The results show a number of differences in the distributional properties of vertical scales derived from IRT models applied at the item and testlet levels. (Contains 4 figures, 8 tables, and 53 references.) (Author/SLD)

RUNNING HEAD: Vertical Scaling with Testlets

Comparing Vertical Scales Derived from Dichotomous and Polytomous IRT Models for a Test Composed of Testlets

N. Scott Bishop

Md Hafidz Omar

Riverside Publishing

Paper presented at the Annual Meeting of the

National Council on Measurement in Education

New Orleans

April, 2002

Abstract

Previous research has shown that testlet structures often violate important assumptions of dichotomous IRT (D-IRT) models—applied to item-level scores—that can in turn affect the results of many measurement applications. In this situation, polytomous IRT (P-IRT) models—applied to testlet-level scores—have been used as an alternative. The objective of this study was to examine the distributional characteristics of vertical scales created using selected D-IRT and P-IRT models for a test composed of testlets. *ITBS* Reading Comprehension test scores from 60,000 randomly selected students (10,000 students per grade from Grade 3 through Grade 8) who took the *ITBS* during a recent fall administration were used in this study. For this data, vertical scales were produced using three D-IRT models—the Rasch model (RM) and the three-parameter logistic (3PL) model calibrated under concurrent and separate group options—and four P-IRT models—the nominal model (NM), the graded response model (GRM), the partial credit model (PCM) and the generalized partial credit model (GPCM). The results showed a number of differences in the distributional properties of vertical scales derived from IRT models applied at the item and testlet levels.

Comparing Vertical Scales Derived from Dichotomous and Polytomous IRT Models for a Test Composed of Testlets

Nearly all multilevel achievement tests utilize developmental score scales. These scales serve a significant role in educational measurement. For test developers, developmental score scales are often the primary scale for multilevel tests. That is, they are the scales from which other auxiliary scales are derived. For test users, developmental score scales are required for examining achievement growth patterns of students across grade levels. This task represents one of the most important uses of standardized achievement test scores. Clearly, the processes used to derive vertical scales, referred to as "scaling to achieve comparability" in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999), represent a significant application of current psychometric theories and techniques. Consequently, the methods used to derive such scales, and in particular the resulting scale characteristics, should be studied.

Although a variety of procedures exist for creating vertical scales, this paper focuses only on methods based on the application of Item Response Theory (IRT). Two IRT models are frequently used for vertical scaling: the one-parameter logistic (1PL) model (see Gardner, Madden, Rudman, Karlsen, Merwin, Callis, & Collins, 1985) and the three-parameter logistic (3PL) model (see CTB, 1984, 1989). Early applications of these models yielded growth trends that were different from other scaling methods (Anastasi, 1958; Phillips & Clarizio, 1984; Hoover, 1984; Burket, 1984; Yen, 1986; Clemans, 1993). A common perception of IRT derived vertical scales is that they suggest lower achieving students grow more rapidly than higher achieving students. As a result, scale variability tends to decrease across grade levels (Hoover,

1984; Yen, 1986). At times, the Rasch model (RM) can also produce scales with these characteristics. For example, the Stanford Achievement Test (SAT), which was scaled using the RM, shows this pattern in its expanded score scale (Gardner et al., 1985). This pattern of decreasing scale variances within grades, from fall to spring, and across grades has been termed "scale shrinkage" (Camilli, 1988).

Many authors, such as Yen (1983, 1985, 1986, 1988), Hoover (1984a, 1984b, 1988), Camilli (1988), and Camilli, Yamamoto, and Wang (1993) have attempted to explain why IRT vertical scales have these characteristics. Hoover (1984a) claimed that scale shrinkage is unfounded empirically as well as intuitively and suggested that this pattern is an artifact of IRT scaling methodologies. Using simulated data, Yen (1986) showed how the treatment of a multidimensional space as a unidimensional ability vector could cause this pattern. Camilli (1988) suggested that scale shrinkage could be the product of systematic estimation errors, measurement errors, and unobtainable ability estimates that might result from a mismatch between item difficulty and examine ability. For example, in two simulated data sets that differed in the variability of the item difficulties, relatively more scale shrinkage was observed when the variability of the item difficulties was smaller. Camilli (1988) and Camilli et al. (1993) found that for groups with large differences in average ability, the one with the smaller amount of measurement error had less variability in their ability scores. For the 3PL model, it is generally agreed that there is greater measurement error for groups at the lower end of the ability continuum than for groups at the higher end.

Although previous vertical scaling studies have used testlet-based data, the effects of the testlet structures have never been a research focus. Consequently, little is known about the psychometric properties of vertical scales for tests composed of testlets. Lee, Brennan, and

Frisbie (2000) defined a testlet as a "subset of items in a test form that is treated as a measurement unit in test construction, administration, and/or scoring." An example of a testlet-based test is a passage-based reading comprehension test. Here, each reading selection has an associated set of questions about it. Because all items in a particular subset pertain to a specific passage selection, they represent a stimulus-based testlet. There has been an increased interest in the psychometric properties of testlets. Many large-scale assessment programs now employ testlet structures (in the form of content-dependent item sets, item bundles, etc.) within their content area tests. Some tests are composed exclusively of testlets.

Despite their widespread usage, it is acknowledged that testlet structures can violate the assumptions of dichotomous IRT (D-IRT) models, particularly local item independence, when applied to item-level scores. Local item independence means that the responses of examinees to any pair of items are statistically independent when the abilities underlying test performance are controlled. Under these conditions, the probability of a response pattern over a set of items equals the product of the individual item response probabilities (Hambleton, 1989; Hambleton, Swaminathan, & Rogers, 1991). A logical question to ask is whether violation of the local independence assumption might also be a factor related to scale shrinkage.

Many vertical scaling studies employ D-IRT methodologies even though the items may not exhibit the necessary property of local item independence. Although sometimes ignored in practice, violation of this assumption has been shown to have non-trivial consequences for many measurement applications. For example, inflated estimates of score reliability, precision, and test information can occur (Sireci, Wainer, & Thissen, 1991; Thissen, Steinberg, & Mooney, 1889; Wainer, 1995; Wainer & Thissen, 1996; Wainer & Wang, 2001). Thissen et al. (1989)

observed lower validity correlation coefficients between scores from an external criterion and ability scores derived with traditional IRT procedures for a testlet-based test.

Several studies have shown that local item dependence can bias item parameter and ability estimates and that the magnitude of the effect seems related to degree of dependence (Ackerman, 1987; Ackerman & Spray; 1986; Yen, 1993; Wainer & Wang, 2001). In a recent study, Wainer and Wang (2001) showed overestimates in guessing parameters occurred for both reading comprehension and listening comprehension items when testlet-associated local item dependence was ignored. Underestimates in item discrimination parameters occurred for listening comprehension items while overestimates occurred for reading comprehension items. Finally, although Huynh (1994) has shown that the Rasch and partial credit model yield the same results under some conditions, when local item dependence exists, the two models can produce different results (see Wilson, 1988). These findings highlight the importance of investigating the properties of vertical scales for testlet-based tests when methods acknowledging the test structure are properly applied.

Although the local item independence assumptions might not hold at the item level, it often can be met for the testlet-level scores. In fact, applying polytomous IRT (P-IRT) models to testlet-level scores has been suggested as an alternative that might ameliorate these problems in measurement applications such as estimating reliability and conditional standard errors of measurement, horizontal equating, and DIF analysis. (See Lee et al., 2000 for a listing of representative research using testlet scores in these areas.) Given the potential utility of P-IRT models in these areas, in conjunction with the knowledge that testlet associated local item dependence has been shown to affect other psychometric applications, an investigation of the properties of P-IRT derived vertical scales for testlet-based tests seems warranted.

A search of the relevant literature did not uncover any prior research comparing D-IRT and P-IRT derived vertical scales for testlet-based tests. However, a recent study conducted by Lee, Kolen, Frisbie, and Ankenmann (2001) compared horizontal equating results derived from P-IRT models (applied to testlet-level scores) to those derived from D-IRT (applied to item-level scores) and classical equating procedures. Using a random-equivalent-groups design, these authors compared observed- and true-score equating relationships from 3PL, graded response, and nominal models using results from traditional equating methods as baselines. Alternate forms for three *ITBS* tests that employed testlet structures were equated: Reading Comprehension, Maps and Diagrams, and Math Problem Solving and Data Interpretation. Noteworthy differences (nearly two raw-score points at some locations on the raw score scale) were observed between the equated scores derived from P-IRT and D-IRT models. The authors noted that differences tended to be more pronounced at the lower and upper ends of the score scale and for tests that most strongly violated the IRT assumptions.

Based on Lee et al's findings, one might speculate that vertical scales derived using testlet-level procedures might exhibit differences from those derived using item-level procedures at the lower and upper grades where previous trends in means and variances have been observed. However, one should be cautious about generalizing horizontal equating results to the vertical scaling scenario. Vertical scaling differs from horizontal equating in several ways. First, in contrast to horizontal equating, alignment of the content specifications and statistical properties between forms is not a paramount issue in vertical scaling. In fact, ascending test levels have progressively increasing item difficulties. In addition, the dimensionality across ascending test levels is in question. For these reasons, Skaggs and Lissitz (1986) distinguished these two tasks in their review of IRT equating research.

Lee et al. (2000) noted that the most appropriate analyses for some tests would be carried out using testlets explicitly. However, the authors acknowledged that testlets might be ignored if it can be demonstrated that the testlets have no practical consequences for the measurement issue in question. Toward this end, the purpose of this study is to explore the effects of testlets on vertical scaling results by examining the distributional properties of vertical scales created using D-IRT models (applied at the item level) and P-IRT models (applied at the testlet level).

Method

Instrument

Data from Form K of the *Iowa Tests of Basic Skills* (*ITBS*) Reading Comprehension test (Hoover, Hieronymus, Frisbie, and Dunbar, 1994) was used in this study. Like most passage-based reading comprehension tests, the *ITBS* Reading Comprehension test is comprised entirely of testlets. Consequently, one might expect that the assumption of local item independence would not hold for the within-testlet items. Table 1 presents the following characteristics for the tests across levels 9 – 14 (Grades 3 – 8): the number of individual items, the number of testlets (passage/item sets), and the number of items within each testlet.

The *ITBS* Reading Comprehension test overlaps test items and passages across adjoining test levels. This overlap, in terms of the number individual items and testlets, is also presented in Table 1. The overlapping items and testlets were treated as "common items" in the vertical scaling procedures described below. Testlet scores were calculated by summing the item-level scores for all items nested within a passage.

Subjects

ITBS Reading Comprehension test scores from 60,000 randomly selected Iowa students (10,000 students per grade from Grade 3 through Grade 8) who took the *ITBS* (Form K) during a

recent fall administration were used in this study. Because of the cross-sectional nature of this design, where data at only one time during the school year was used, only across-grade scale properties (shrinkage) are addressed in this paper.

IRT Assumptions

Exploratory factor analysis (EFA) was used to investigate the dimensionality of the item-level and testlet-level scores. All analyses were conducted using Mplus (Muthen & Muthen, 1998). Both the item-level and testlet-level scores were treated as categorical variables. For categorical variables, Mplus estimates latent means and threshold structures in addition to factor loadings and latent covariance structures. Muthen (1984) has shown similarities between his factorial model with threshold structures and IRT models. Specifically, the threshold structures are like IRT difficulty-parameter estimates while the factor loadings are like IRT slope-parameter estimates. Because the dependent variables were categorical, the usual assumption of multivariate normality between explanatory variables was relaxed and no distributional assumptions were placed on the latent factors. This is also consistent with IRT where the items are not considered continuous (let alone normally distributed) and the latent dimensions are typically continuous. A weighted-least-squares estimation procedure was employed.

Yen's Q_3 statistic, between and within testlets, was used to assess local item independence for the D-IRT models. The computer program IRTNEW (Chen, 1998) was used to compute Q_3 for each item pair. For the purposes of this study, only local item independence relative to the 3PL D-IRT was investigated. This was because testlet-level analysis is a recommended strategy when local dependence is observed at the item level (Ferrara et al., 1997; Yen, 1993). Q_3 is computed as follows. First, expected examinee performance on items is

determined using available ability and item-parameter estimates. Next, deviations (residuals) between the examinees' expected and observed performance is determined for each item. Finally, the value of Q_3 for an item pair is simply the correlation between their respective deviations. According to Yen (1993), the expected value for Q_3 is $-1/(k\text{-}1)$, where k equals the number of test items. Yen notes that the expectation is slightly negative because item scores are involved in the calculation of the ability estimates, which constitutes a part-whole contamination. For a test with k items, there are a total of $k(k\text{-}1)/2$ item pairs for which Q_3 can be computed.

Design

This study compares the distributional characteristics of vertical scales derived under three D-IRT models—the Rasch model (RM) and the 3PL model calibration with concurrent group and separate group options—and four P-IRT models—the partial credit model (PCM), the generalized partial credit model (GPCM), the graded-response model (GRM), and the nominal model (NM). These models were selected for the following reasons. First, the Rasch and 3PL models are frequently used in practice. Second, unlike the separate-group calibration, the 3PL model concurrently calibrated with the group option provides an explicit opportunity to estimate the first two latent moments of the ability distributions during marginal-maximum-likelihood estimation (Bock & Zimowski, 1997). Third, the selected P-IRT procedures cover many of the major polytomous models. This allows for a number of interesting comparisons. For example, the two different conceptualizations of the PC model can be compared (where the GPCM relaxes the assumption of a common discrimination parameter across all items). Additionally, the PCM, GPCM, and GRM assume that there is an ordered quality to the testlet scores whereas the NM makes no such assumption.

The GPCM and GRM extend the two-parameter D-IRT model for use with polytomous items. That is, both models do not include a "guessing" parameter and allow the discrimination parameter to vary across items. The two models develop their probability functions differently, however. The GPCM, credited to Muraki (1992, 1997), is very similar to Master's PCM (Masters, 1982). However, Master's model does not allow its discrimination parameter to vary. In the GPCM, the probability that an examinee with ability theta will score k on item j is expressed as:

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=0}^{k} Da_j(\theta - b_{jv})\right]}{\sum_{c=0}^{m_j} \exp\left[\sum_{v=0}^{c} Da_j(\theta - b_{jv})\right]}$$

In this model, j is the item, m_j is the number of response categories for item j, and k is the response category of interest. D represents the scaling constant (where D=1.7 is used for the normal ogive approximation), a_j is the slope parameter for each response category and b_{jv} is the item category parameter. The a_j parameter governs the spread of the step curve while b_{jv} indicates the location where two adjacent categories have equal probabilities.

In contrast, the GRM, credited to Samejima (1969, 1972, 1997), models the category k probability by using a series of dichotomous two-parameter functions. Specifically,

$$P_{jk}(\theta) = \frac{1}{1+\exp\left[-Da_j(\theta - b_{jk})\right]} - \frac{1}{1+\exp\left[-Da_j(\theta - b_{j(k+1)})\right]}$$

where D is the scaling constant, a_j is the slope parameter for item j, and b_{jk} is the item-category difficulty parameter. The value of b_{jk} is the point on the theta scale at which the first ratio in the

equation passes 50% of the responses in category k+1 or higher. When k=0, the first ratio in the equation is evaluated to be 1. When $k=m_j+1$, the second ratio becomes 0.

Unlike the previous models, the NM, credited to Bock (1972), does not assume ordinality of the category difficulty parameters. The NM is mathematically operationalized in the following category k response function:

$$P_{jk}(\theta)=\frac{\exp\left[a_{jk}\theta+c_{jk}\right]}{\sum_{v=0}^{m_j}\exp\left[a_{jv}\theta+c_{jv}\right]}$$

where a_{jk} is the slope parameter for category k of item j, and c_{jk} is the intercept parameter for the nominal category k of item j. The slope parameter governs the spread of the category response curve while the intercept parameter indicates the item-category difficulty interaction with the slope parameter.

BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used for scaling under the 3PL model both for the concurrent calibrations with groups and the separate calibrations without groups. For the separate groups calibration, the ST computer program (Hanson & Zeng, 1995) was applied to the common-item parameters obtained from BILOG-MG to determine the scaling constants used to derive the vertical scale. The 3PL concurrent calibration with groups required no scaling constants. WINSTEPS (Linacre & Wright, 1998) was used to derive item and ability estimates for the RM and the PCM. MULTILOG (Thissen, 1991) was used to estimate the item parameter and ability estimates used for constructing vertical scales under the NM and GRM, while PARSCALE (Muraki & Bock, 1997) was used for the GPCM.

While WINSTEPS uses the unconditional maximum-likelihood-estimation algorithm, BILOG-MG, PARSCALE, and MULTILOG all implement marginal-maximum-likelihood

estimation algorithms. In general, defaults were used for most program options. However, the "repeat" option was employed on PARSCALE's Block card to allow each testlet to have its own step value. Also, 20 quadrature points were used for parameter estimation when programs allowed for user manipulation of this feature. Maximum-likelihood estimation was used to estimate theta scores for all models. In order to facilitate comparisons among all the vertical scales, a common metric was required. This was accomplished by scaling all Grade 3 ability distributions to have a mean of zero and a standard deviation of one.

The version of MULTILOG used for this study only accommodated items with 10 categories. However, one testlet at Grade 8 consisted of 11 categories. To address the program requirements, the two upper score points for this testlet were collapsed into one category (as these two scores affected the smallest number of examinees). It is not believed this poses an obstacle. In fact, such collapsing can be advantageous during estimation when categories have small counts (Wainer, Sireci, & Thissen, 1991). This procedure does not affect the other categories as they would retain their response frequencies. Although the GRM and NM vertical scales were derived differently than the PCM and GPCM, this only affects the Grade 8 results.

For the RM, PCM and GPCM, scaling constants were derived using the mean/mean method. In the case of the RM and PCM, this essentially entailed determining the difference in the item/step difficulties among the common items—as described in Masters (1984) but without differentially weighting more precisely estimated steps. Estimation of the scaling coefficients for the NM and GRM was accomplished with a polytomous extension of the test characteristic curve approach as developed by Baker (1992 and 1993b) and implemented with the EQUATE program (Baker, 1993a). An additional program issue arose because EQUATE only allows for items with nine categories. Therefore, another practical decision was required for managing this

situation. Specifically, one common testlet (consisting of 10 categories) between the Grade 7 and 8 tests was not used in the determination of the scaling constants. Scaling constants were estimated on the remaining three common testlets, which consisted of 18 total categories. This situation might have been managed in other ways (e.g., collapsing the 10 category testlet into 9 categories), which in turn would have likely resulted in different scaling constants. As before, this issue only impacts the Grade 8 results for the GRM and the NM.

Results

Descriptive Statistics

Descriptive statistics (including means, standard deviations, skewness and kurtosis indices, and KR-20 reliabilities) for the tests at each grade level are reported in Table 2.

Unidimensionality Assumption

The EFA results are provided in Table 3, which reports the eigenvalues and the root mean squares (RMS) for the off-diagonal residuals for several models that differ in the number of factors. Regarding item-level scores, there were a number of eigenvalues greater than one. At all test levels, the RMS residual for the first factor on the item-level scores is greater than 0.05. Additionally, visual inspection of the resulting pattern matrices (Promax rotation) revealed that items nested within the same testlet tended to load on the same factor. On the other hand, there seemed to be a dominant factor relative to Reckase's (1979) recommendation that the first factor account for a minimum of 20% of the test variance. For testlet-level scores, a one-factor model appears sufficient in all cases.

Local Independence Assumption

Table 4 reports the distributional characteristics of Q_3 for item pairs categorized as between-testlet items and within-testlet items. At all test levels, the mean value of Q_3 for items

between testlets is very close to the expected value. In contrast, the mean value of Q_3 for items within testlets is nearly one standard deviation greater than the expected value at all grade levels. In Figure 1, boxplots are used to further illustrate the magnitude of these differences. At all grades, P_{75} of the between-testlet Q_3 distribution was less than P_{25} of the within-testlet Q_3 distribution. As expected, these results suggest that the assumption of local independence does not hold for the item-level data. This in turn supports the proposed testlet-level analyses.

Vertical Scale Characteristics

Vertical scales based on the theta estimates from seven IRT models were created. Due to the large number of pair-wise comparisons (21), attention will only be given to several selected contrasts. As noted earlier, the RM and PCM were derived using WINSTEPS, which employs a different algorithm than the other programs used in this study. Also, both models are solely functions of the difficulty of the items/steps and result in a one-to-one correspondence between the raw scores and theta estimates (i.e., there is no pattern scoring). For these reasons, it is sensible to contrast the results from these two models apart from the others. All P-IRT theta estimates were derived under separate calibrations; therefore, the NM, GRM, and GPCM results will primarily be compared to the 3PL separate calibration theta estimates. Finally, results from the separate and group calibrations for the 3PL model will also be compared because of the asserted advantages of group calibrations in the vertical scaling scenario.

Table 5 presents the correlations among the theta estimates both within and across grades. The theta estimates for all models tended to be very highly correlated, with nearly all correlations being greater than 0.94 in magnitude. The RM and PCM yielded the greatest correlations, 0.996 and above in all cases. The 3PL separate and group theta estimates were also highly correlated, but to a slightly less degree (typically around 0.990). Correlations between the 3PL separate

calibration and P-IRT theta estimates varied somewhat. Although many of these correlations were high in magnitude, they were typically lower than the correlations among the various P-IRT theta estimates.

At Grade 8, the correlations involving the 3PL group theta estimates were noticeably lower than at the other grades. As evidenced in Table 6, there was a ceiling effect at this grade for the theta estimates derived from Phase 3 of the BILOG-MG group calibration (specifically, P_{95} and P_{99} have the same value). This occurred because a number of these theta estimates were assigned a common maximum value. The magnitude of correlations among the Grade 8 theta estimates for this model were likely affected, in part, by this ceiling effect.

Inspection of the within-grade scatterplots for the theta estimates revealed slight, yet noticeable, non-linear trends. Figure 2 presents the Grade 3 scatterplots. (Scatterplots for the other grades are not presented as they were very similar to those at Grade 3.) The scatterplots involving the 3PL separate and group theta estimates tended to exhibit a "dog-leg" pattern at the lower end of the theta scale. Several of the theta estimates from the P-IRT models also showed curvilinear trends, both at the lower and upper theta values. These patterns were observed at all grades. Finally, the correlations among the theta estimates across all grade levels were often lower than their within grade counterparts. Although across-grade correlations are often greater in magnitude than within-grade correlations because of increased variability, the cumulative effect of the nonlinear trends observed in the within-grade scatterplots likely lowered the across-grade correlations in some instances. Scatterplots for the estimated thetas across all grades are also presented in Figure 2.

Table 6, in conjunction with Tables 7 and 8 as well as Figures 3 and 4, provide information regarding the distributional characteristics of the theta estimates. Table 6 reports a

number of descriptive statistics for the theta estimates. Figures 3 and 4 graphically summarize many of the more noteworthy trends. As in previous scale shrinkage studies, these figures reflect across grade changes both in the magnitude and variability of the estimated thetas. Figure 3 depicts the grade-to-grade change in the theta estimates at various percentiles (the actual magnitudes of the grade-to-grade changes are listed in Table 7). Yen (1986) has suggested such data needs to be interpreted cautiously because true-score growth of the same amount will not maintain the same percentiles in the observed score distribution across grades because of measurement error. The plots are used here only as descriptive information about the distributions of the theta estimates.

As seen in Figure 3 and documented in Table 7, the theta estimates from the 3PL separate calibrations are fairly typical of previous trends reported for IRT vertical scales. That is, changes at the lower percentiles are greater than the changes at the higher percentiles. The changes in estimated theta for nearly all percentiles for this model became smaller across grades. This general trend was often true for other models as well. Exceptions to this trend are presented in bold type in Table 7. Although the 3PL group theta estimates also show this trend, it is not nearly as pronounced as in the 3PL separate calibrations plot. It is also less pronounced in the GRM and GPCM plots. The PCM, RM, and NM seem least similar to the separate 3PL plot in this respect.

Although plots like those in Figure 3 have been frequently used to illustrate the results from vertical scaling studies, they don't readily facilitate comparisons across models. To better illustrate model differences, Figure 4 provides plots for the mean, standard deviation (SD), P_{10}, and P_{90} trends for all models. For the mean and SD plots, the 3PL group Phase 2 results are also included. These means and SDs (reported separately in Table 9) were directly estimated by the

MMLE algorithm during estimation of item parameters (Phase 2). These directly estimated group parameters, especially the variability indices, are frequently recommended over indices based on theta estimates—i.e., BILOG MG's Phase 3 estimates (Mislevy, 1984; Camilli, 1988; Bock & Zimowski, 1997). Without quantifying the different measurement error variances at each grade level, the true variability at each level may be confounded, thus, complicating inferences about growth trends (Camilli, Yamamoto, & Wang, 1993).

The plots in Figure 4 show that model differences tend to become more pronounced for higher percentiles and across increasing grade levels. Like the percentile trends noted in Figure 3, decelerating changes in the values for the mean, P_{10}, and P_{90} across grades was a common feature for most models. As noted earlier exceptions are presented in bold type in Table 7. The magnitude of the change was considerably different across models. For example, the NM, GRM, and GPCM estimated thetas increased in value more than the 3PL separate calibration estimated thetas. These plots also revealed a great deal of similarity between the RM and the PCM.

The differences in the variability of the theta estimates across grades are also noteworthy. The SD plots clearly show a reduction in variability across grades for most models, but in particular for the 3PL separate calibrations. The magnitude of the SDs across grades are reported in Table 8. SDs that do not decrease across grades are presented in bold type. The SDs for the group 3PL, GRM, and GPCM also tended to decrease across grades. However, the magnitude of the SD reduction relative to their Grade 3 values is not nearly as extreme as in the 3PL separate calibration results. The same trend is also seen for the group 3PL Phase 2 direct SD estimates, a noteworthy finding for testlet-based tests. As depicted in Figure 3, the difference between SDs for the direct SD estimates and the SDs based on theta estimates were fairly constant across grades. The RM and PCM were very similar to each other in terms of SD change across grade

levels. In fact, the SDs for these models were generally close to their initial Grade 3 values. The NM standard deviations also tended to remain close to the Grade 3 values. For the P-IRT models that included multiple discrimination parameters (in the NM case each category had unique discrimination and item-category difficulty parameters), this was the only model that did not reveal at least some reduced variability at the upper grades.

Table 8 also documents the variability trends in terms of interquartile range ($(P_{75} - P_{25})/2$) differences and the $(P_{95} - P_5)/2$ differences. These results generally agree with the trends in the SD plots. Namely, there were marked reductions in these differences for the 3PL separate calibrations across grades. The group 3PL, GRM, GPCM also showed reductions, but not nearly as much as observed for the separate 3PL results. Also, the PCM, RM, and NM differences remained close to their Grade 3 values.

Discussion

To date there has been little or no direct research regarding how testlet structures affect the properties of vertical scales. For this reason it is believed that this study has addressed a gap in current educational measurement research and should contribute to both the testlet and vertical scaling literature. Moreover, this research should also inform practice as the current study involved several commonly used methods for producing vertical scales.

Understanding the properties of vertical scales derived from different methods is important for several reasons. First, vertical scales are often the primary score scale for many multilevel achievement tests (i.e., they are the scales from which other auxiliary scales are derived). Second, vertical scales are the basis upon which interpretations regarding growth are made. Growth is clearly an important educational outcome, particularly for the school improvement movement. Finally, when different methods produce different results, test

developers must make choices about the types of scales they provide to assess growth and evaluate educational outcomes (Becker & Forsyth, 1992). Such choices are nontrivial, particularly for the users of standardized achievement tests who must help students and parents interpret test results.

Although multidimensional tests may manifest symptoms of local dependence, symptoms of local dependence do not necessarily imply that the ability underlying a test is multidimensional. For the Reading Comprehension test used in this study, a factor analysis of the testlet level scores revealed one clear dominant factor. However, the item-level scores, clearly violated local item independence assumption. One might speculate violation of this assumption was at least partially responsibly for the observed differences between the D-IRT and P-IRT derived scales. For this testlet-based data, vertical scales derived from both the item and testlet models tended to exhibit decelerating mean growth and reduced variability (scale shrinkage) across grades. However, the magnitude of the mean growth and variability differences varied considerably across models. Scale shrinkage also occurred when the scales were derived with recommended direct estimates of variability, namely the multiple groups item-level approach. For item-level models, the Rasch model appeared least prone to this trend. This is consistent with what has been observed in past literature (Loyd and Hoover, 1980).

Polychotomous models are often recommended as an alternative IRT approach when the local item independence assumption is violated. These models tended to show considerably less scale shrinkage, especially compared to the item-level 3PL separate calibration results. In particular, the partial credit model and Bock's nominal model appeared most immune to this pattern. These two models are either less parameterized (the partial credit) or the more parameterized (the Nominal model) than the other polychotomous models considered. The

Nominal model, in particular, stood out from the rest of the models because it does not assume that the categorical responses are ordered (Baker, 1993). The suitability of this assumption for achievement test data may be debatable. Additionally, this model has as many discrimination parameters as category intercept parameters. Within the Rasch family, the partial credit and Rasch model results were very similar. Both these models are solely functions of item/step difficulties. Also, the raw-scores provide sufficient statistics for deriving the theta estimates.

A strength of this study is that several competing models were used to develop vertical scales from the same test data. Also, the available sample size at each grade level was well above the minimum generally recommended for IRT calibrations. Finally, the *ITBS* Reading Comprehension test is similar to other passage-based reading comprehension tests. Because it is recognized that items attached to the same reading passage can cause local item dependence, one would expect that these results might generalize to other tests of this type.

Because research regarding the psychometric properties of testlets has not adequately addressed the subject of vertical scaling to date, issues related to creating vertical scales for testlet-based tests should be a promising area for future research. Indeed, the current study has several limitations that might be addressed by future studies. For example, these analyses were based on data from only one kind of test (a passage-based reading comprehension) with the focus on Grades 3 – 8. To establish the generalizability of these results it would be necessary to 1) conduct additional studies using other content domain tests and grade levels, and 2) create simulated data sets that control for the magnitude of item dependence and dimensionality. Additionally, this study only focused on the distributional characteristics of the vertical scales. Other properties and outcomes, such as model-data fit and information as well as within-grade shrinkage, should also be considered. Based on the findings from prior testlet research, one

might expected that future studies will show that differences depend on the degree to which the testlet structure violates these assumptions.

Other IRT models could be employed in this situation. Even those who have recommended applying P-IRT models to score testlets recognize that this approach has limitations. For example, this method might lose information that the item-level pattern holds. Such procedures might also be difficult to apply in some contexts, like adaptive testing. In response to these concerns, Wainer and Wang (2001) have studied a new testlet-level model that is applied to the item-level scores. Specifically, this model extends the traditional 3PL model by including a random effect component for the person-by-testlet interaction.

Within IRT methodologies, a wide variety of procedural and design options exist. For example, this study utilized common items across test levels to determine the scaling constants used to create the vertical scales. Scaling tests consisting of an external set of common items have been employed in other situations. Also, common-person designs have been employed (e.g., where examinees take two adjoining levels). In this study all passage-related items were included in the testlet scores. However, Yen (1993) recommended that testlets only be created for items that exhibited dependence. These are important design and procedural variations that might affect the characteristics of the P-IRT derived vertical scales. Consequently, the effects of these variations should be addressed in future studies.

A number of practical decisions were applied in this study because of software limitations (e.g., collapsing categories). Both the program limitations and the resulting practical decisions noted earlier would need to be addressed if P-IRT procedures are applied in practice. For example, no reference could be found regarding the application of a test-characteristic-curve procedure to derive scaling constants for the GPCM.

References

Ackerman, T. A. (1987, April). The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence. Paper presented at the annual meeting of the American Educational Research association, Washington, DC.

Ackerman, T. A., & Spray, J. A. (1986, April). A general model for item dependence. Paper presented at the annual meeting of the American Educational Research association, San Francisco.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

Anastasi, A. (1958). Differential Psychology (3rd ed.). New York: Macmillan.

Baker, F.B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, 16, 87-96.

Baker, F.B. (1993a). Equate 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement*, 17, 20.

Baker, F.B. (1993b). Equating tests under the nominal response model. *Applied Psychological Measurement*, 17, 239-251.

Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement*, 29(4) 341-354.

Bock, R. D. & Zimowski, M. F. (1997). Multiple Groups IRT. In *Handbook of Modern Item Response Theory*, 433-448.

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practices*, 3(4), 15-16.

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics, 13*(3), 227-241.

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement*, 36(1), 73-78.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, 17(4), 379-388.

Chen, W. H. (1998) IRTNEW [Computer Program]. Available at www.unc.edu/~dthissen/dl.html.

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational assessment*, 1(4), 329-347.

CTB/McGraw-Hill. (1974). *Comprehensive Tests of Basic Skills, Form 5: Technical Bulletin No. 1*. Monterey, CA: Author.

CTB/McGraw-Hill. (1984). *Comprehensive Tests of Basic Skills, Form U and V: Technical Report*. Monterey, CA: Author.

CTB/McGraw-Hill. (1989). *Comprehensive Tests of Basic Skills, Fourth Edition: Technical Bulletin No. 1*. Monterey, CA: Author.

Ferrara, S., Huynh, H., and Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education*, 10(2), p. 123 –146.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (ed.), *Educational measurement* (3rd ed., pp. 143-200). New York, NY: Macmillan.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hanson, B. & Zeng, L. (1995). ST: A Computer Program for IRT Scale Transformation: Version 1.0. Available at www.uiowa.edu/~itp/pages/swibm.shtml.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1994). *Iowa Tests of Basic Skills, Forms K and L, interpretive guide for school administrators*. Chicago: Riverside Publishing Co.

Huynh, H. (1994). On the equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometricka,* 59, 111-119.

Lee, G., Brennan, R., & Frisbie, D. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practices, 19*(1), 9 – 15.

Lee, G., Kolen, M. J., Frisbie, D.A., & Ankenmann, R.D. (2001). A comparison of the performance of dichotomous and polytomous IRT models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*(4), 357-372.

Linacre, J. M. & Wright, B.D. (1998). *WINSTEPS [Computer Program]*. Chicago, IL: Mesa Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Masters, G. N. (1984). Constructing an item bank using partial credit scoring. *Journal of Educational Measurement*, 21, 19-32.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement*, 16, 159-176.

Muraki, E. & Bock, R.D. (1997) *PARSCALE version 3: IRT Item Analysis and Test Scoring for Rating Scale Data [Computer Program]*. Chicago, IL: Scientific Software.

Muthen, B. O. (1984) A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.

Muthen, L. K. & Muthen, B, O. (1998) . Mplus: A Statistical Analysis With Latent Variables [Computer Program]. Los Angeles, CA: Muthen & Muthen.

Philips, S.E. & Clarizio, H. F. (1988). Conflicting Growth Expectations Cannot Both be Real: A Rejoinder to Yen. *Educational Measurement: Issues and Practices*, 7(4), 18-19.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics*, 4, 207-230.

Schulz, E. M., & Nicewander, W. A., (1997, May). Grade equivalent and IRT representations of growth. *ACT research report* 97-2.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: the sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and policy analysis*, 16(1), 41-49.

Sireci, S. G., Wainer, H., & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of educational measurement*, 28, 237-247.

Skaggs, G. & Lissitz, R. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, 56(4), 495-529.

Stocking, M. L. & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, 7, 201-210.

Thissen, D. (1991). *MULTILOG [Computer program]*. Chicago, IL: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement*, 26(3), 247-260.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based tests: The 1991 Law School Admissions Tests as an example. *Applied Measurement in Education*, 8(2), 157-187.

Wainer, H. Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of educational measurement,* 28(3), 197-219.

Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational measurement: Issues and practices*, 15(1), 22-29.

Wainer, H. & Wang, X. (2001, May). Using a new statistical model for testlets to score TOEFL. *TOEFL Technical Report* TR-16. Educational Testing Service.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, 12, 353-364.

Yen, W. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, 50, 399-410.

Yen, W. M. (1986). The Choice of Scale for Educational Measurement: An IRT Perspective. *Journal of Educational Measurement*, 23(4), 299-325.

Yen, W. M. (1988). Normative Growth Expectations Must Be Realistic: A Response to Phillips and Clarizio. *Educational Measurement: Issues and Practices*, 7(4), 16-17.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Educational Measurement: Issues and Practices*, 7(4), 16-17.

Yen, W. M., Burket, G. R., & Fitzpatrick, A. R. (1995). Response to Clemans. *Educational Assessment*, 3(2), 181-190.

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R.D. (1996). *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items [Computer program]*. Chicago, IL: Scientific Software.

Table 1. Characteristics of the *ITBS* Reading Comprehension Test at each Grade

Test Characteristic	Grade 3	4	5	6	7	8
Items	36	38	41	44	46	49
Testlets	7	8	8	7	7	8
Items in each Testlet	6,4,6, 6,5,3,6	6,5,3,6, 5,4,3,6	5,4,3,6, 8,4,5,6	8,4,5,6, 7,7,7	7,7,7, 9,4,7,5	9,4,7,5, 5,6,3,10
Item overlap Below	NA	20	18	23	21	25
Testlet overlap Below	NA	4	4	4	3	4
Item overlap Above	20	18	23	21	25	NA
Testlet overlap Above	4	4	4	3	4	NA

Table 2. Descriptive Statistics for *ITBS* Reading Comprehension Raw Scores at each Grade

Sample Statistics	Grade					
	3	4	5	6	7	8
n-count	10,000	10,000	10,000	10,000	10,000	10,000
Mean	19.32	21.68	24.65	28.17	27.92	27.74
Median	19.28	21.99	25.25	29.13	28.43	27.75
SD	7.39	7.24	8.05	8.81	8.93	9.78
Skewness	0.00	-0.12	-0.23	-0.37	-0.20	0.01
Kurtosis	-0.92	-0.80	-0.80	-0.76	-0.82	-0.87
KR-20	0.87	0.86	0.88	0.90	0.89	0.90
N Max	6	9	14	20	18	12/24

Notes. N Max documents the number of students with perfect test scores. At grade 8, 12 examinees had perfect raw scores. However, 24 perfect scores resulted when the top two score levels were collapsed in order to accommodate MULTILOG.

Table 3. Exploratory Factor Analysis Results for Item- and Testlet-Level Scores

	Grade Level											
	3		4		5		6		7		8	
Factor	Eigen	RMS	Eigen	RMS	Eigen	RMS	Eigen	RMS	Eigen	RMS	Eigen	RMS
	Item Level Scores											
1	10.57	0.065	10.35	0.080	11.66	0.087	14.23	0.084	13.77	0.063	14.03	0.080
2	1.79	0.044	1.75	0.062	1.91	0.062	2.11	0.061	1.98	0.053	2.00	0.065
3	1.29	0.037	1.49	0.046	1.31	0.048	1.28	0.046	1.34	0.043	1.44	0.054
4	1.12	0.031	1.25	0.035	1.25	0.036	1.14	0.038	1.22	0.038	1.33	0.044
5	1.07	0.027	1.09	0.030	1.14	0.028	1.07	0.034	1.12	0.034	1.16	0.039
6	1.00	0.024	1.06	0.025	1.07	0.024	1.04	0.028	1.06	0.030	1.06	0.034
7	0.97	0.021	1.02	0.022	1.03	0.020	0.98	0.025	1.05	0.026	1.00	0.031
8	0.94	0.018	0.99	0.020	0.95	0.019	0.94	0.022	0.95	0.023	0.94	HW
	Testlet Level Scores											
1	3.59	0.036	3.73	0.028	3.94	0.034	3.99	0.039	3.86	0.041	4.38	0.028
2	0.89	0.008	0.88	0.008	0.74	0.008	0.69	0.009	0.76	0.009	0.72	0.015
3	0.62	0.003	0.72	HW	0.66	0.007	0.65	0.003	0.68	0.005	0.63	0.006
4	0.53		0.60		0.64	0.002	0.49		0.51		0.56	0.002
5	0.50		0.56		0.62		0.43		0.44		0.50	
6	0.45		0.53		0.57		0.41		0.39		0.45	
7	0.42		0.51		0.45		0.34		0.36		0.40	
8			0.45		0.39						0.36	

Notes. Eigen = Eigenvalue; RMS = Root Mean Square of the off-diagonal residuals; HW = Heywood Case. All analysis conducted with Mplus using a weighted-least-squares estimation procedure.

Table 4. Distribution of Q_3 Between and Within Testlets at each Grade

Level	No. of Q_3	E(Q_3)	Mean	Diff.	SD	Skew	Kurt	Min - Max
Grade 3	630	-0.0286						
Between	551		-0.0299	0.0013	0.025	-0.332	0.277	-0.119 - 0.040
Within	79		0.0270	0.0556	0.045	1.592	4.719	-0.040 - 0.235
Grade 4	703	-0.0270						
Between	626		-0.0290	0.0020	0.024	-0.167	0.545	-0.109 - 0.067
Within	77		0.0291	0.0561	0.056	2.323	8.644	-0.039 - 0.316
Grade 5	820	-0.0250						
Between	727		-0.0300	0.0050	0.025	0.454	0.036	-0.103 - 0.038
Within	93		0.0375	0.0625	0.058	1.392	1.569	-0.048 - 0.242
Grade 6	946	-0.0233						
Between	824		-0.0280	0.0047	0.021	-0.099	0.368	-0.114 - 0.067
Within	122		0.0332	0.0565	0.046	2.141	8.536	-0.045 - 0.292
Grade 7	1035	-0.0222						
Between	899		-0.0234	0.0012	0.019	-0.203	0.388	-0.095 - 0.051
Within	136		0.0154	0.0376	0.040	0.183	1.124	-0.112 - 0.137
Grade 3	1176	-0.0208						
Between	1030		-0.0229	0.0021	0.019	-0.221	0.181	-0.093 - 0.034
Within	146		0.0240	0.0448	0.041	0.423	3.374	-0.132 - 0.168

Notes. Expected value given by Yen (1993) as $-1/(k-1)$. Number of possible item pairs is $k(k-1)/2$. Figure 3 graphically depicts the between- and within-testlet differences.

Table 5. Within and Across Grade Correlations between Estimated Thetas

Grade 3 Above Diagonal/Grade 4 Below Diagonal

	3PL S	RM	PCM	NM	GRM	GPCM	3PL G
3PL S		0.955	0.951	0.966	0.949	0.949	0.997
RM	0.964		0.999	0.987	0.989	0.992	0.962
PCM	0.960	0.999		0.986	0.987	0.994	0.958
NM	0.972	0.992	0.991		0.985	0.991	0.970
GRM	0.959	0.989	0.986	0.988		0.990	0.956
GPCM	0.960	0.993	0.994	0.995	0.990		0.956
3PL G	0.992	0.971	0.968	0.978	0.964	0.968	

Grade 5 Above Diagonal/Grade 6 Below Diagonal

	3PL S	RM	PCM	NM	GRM	GPCM	3PL G
3PL S		0.972	0.966	0.973	0.969	0.967	0.990
RM	0.972		0.998	0.990	0.992	0.996	0.981
PCM	0.961	0.996		0.990	0.987	0.998	0.976
NM	0.975	0.988	0.987		0.981	0.993	0.977
GRM	0.964	0.988	0.979	0.982		0.987	0.976
GPCM	0.959	0.990	0.994	0.990	0.985		0.976
3PL G	0.989	0.983	0.977	0.985	0.974	0.976	

Grade 7 Above Diagonal/Grade 8 Below Diagonal

	3PL S	RM	PCM	NM	GRM	GPCM	3PL G
3PL S		0.967	0.959	0.976	0.962	0.960	0.989
RM	0.952		0.998	0.992	0.991	0.994	0.972
PCM	0.945	0.999		0.990	0.987	0.996	0.965
NM	0.969	0.987	0.984		0.987	0.992	0.978
GRM	0.950	0.989	0.985	0.987		0.988	0.967
GPCM	0.947	0.995	0.995	0.989	0.991		0.964
3PL G	0.981	0.927	0.918	0.951	0.931	0.919	

Grades 3 - 8

	3PL S	RM	PCM	NM	GRM	GPCM	3PL G
3PL S							
RM	0.941						
PCM	0.938	0.999					
NM	0.940	0.981	0.978				
GRM	0.964	0.988	0.986	0.973			
GPCM	0.952	0.989	0.991	0.963	0.990		
3PL G	0.983	0.974	0.972	0.972	0.979	0.972	

Table 6. Summary Statistics for Estimated Thetas at each Grade

	3PL S	3PL G	RM	PCM	NM	GRM	GPCM
Grade 3							
Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SD	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Skew.	-0.62	-0.52	0.24	0.39	0.09	0.20	0.54
Kurt.	1.13	1.11	-0.07	0.29	0.62	-0.31	0.77
P_{01}	-3.21	-3.23	-2.02	-1.98	-2.36	-2.06	-2.03
P_{05}	-1.79	-1.67	-1.46	-1.42	-1.59	-1.53	-1.46
P_{10}	-1.22	-1.18	-1.31	-1.27	-1.23	-1.27	-1.19
P_{25}	-0.55	-0.56	-0.78	-0.76	-0.66	-0.76	-0.72
P_{50}	0.09	0.07	-0.08	-0.10	0.00	-0.03	-0.07
P_{75}	0.64	0.64	0.65	0.61	0.64	0.72	0.61
P_{90}	1.13	1.17	1.26	1.25	1.23	1.31	1.29
P_{95}	1.45	1.51	1.67	1.70	1.61	1.64	1.73
P_{99}	2.15	2.19	2.65	2.86	2.59	2.39	2.74
Grade 4							
Mean	0.66	0.67	0.73	0.73	0.69	0.77	0.87
SD	0.78	0.86	0.95	0.94	0.95	0.94	0.93
Skew.	-0.53	-0.37	0.23	0.39	0.14	0.09	0.46
Kurt.	1.10	1.35	0.01	0.27	0.50	-0.33	0.59
P_{01}	-1.92	-1.83	-1.17	-1.09	-1.47	-1.19	-0.97
P_{05}	-0.63	-0.72	-0.70	-0.64	-0.82	-0.72	-0.52
P_{10}	-0.28	-0.38	-0.44	-0.40	-0.50	-0.47	-0.28
P_{25}	0.22	0.15	0.04	0.04	0.05	0.06	0.21
P_{50}	0.71	0.69	0.71	0.68	0.69	0.77	0.82
P_{75}	1.17	1.22	1.32	1.29	1.29	1.44	1.46
P_{90}	1.58	1.72	1.93	1.92	1.88	1.99	2.08
P_{95}	1.81	2.00	2.33	2.35	2.22	2.28	2.44
P_{99}	2.32	2.67	3.30	3.43	3.22	2.93	3.40
Grade 5							
Mean	1.12	1.22	1.40	1.43	1.25	1.40	1.71
SD	0.59	0.75	0.99	0.97	0.91	0.83	0.96
Skew.	-0.43	-0.24	0.25	0.49	0.38	0.05	0.53
Kurt.	1.25	1.18	-0.04	0.38	1.19	-0.37	0.70
P_{01}	-0.84	-0.70	-0.69	-0.56	-0.74	-0.35	-0.23
P_{05}	0.13	0.02	-0.12	0.00	-0.20	0.06	0.28
P_{10}	0.42	0.30	0.11	0.22	0.14	0.30	0.55
P_{25}	0.79	0.75	0.74	0.78	0.67	0.80	1.05
P_{50}	1.15	1.23	1.35	1.33	1.23	1.41	1.64
P_{75}	1.49	1.70	2.05	2.02	1.79	1.99	2.28
P_{90}	1.80	2.14	2.70	2.71	2.36	2.48	2.95
P_{95}	2.02	2.43	3.16	3.24	2.74	2.76	3.42
P_{99}	2.52	3.09	3.91	4.11	3.85	3.29	4.44

Table 6. Continued

	3PL S	3PL G	RM	PCM	NM	GRM	GPCM
Grade 6							
Mean	1.38	1.61	2.00	2.04	1.67	1.84	2.31
SD	0.47	0.73	1.08	1.07	1.00	0.76	0.89
Skew.	-0.45	-0.12	0.17	0.51	0.24	0.01	0.54
Kurt.	1.24	1.02	-0.25	0.12	0.58	-0.51	0.35
P_{01}	-0.20	-0.18	-0.29	-0.01	-0.54	0.22	0.59
P_{05}	0.60	0.45	0.35	0.54	0.12	0.60	1.01
P_{10}	0.81	0.72	0.57	0.72	0.44	0.82	1.23
P_{25}	1.11	1.15	1.26	1.29	1.01	1.28	1.67
P_{50}	1.40	1.60	1.96	1.91	1.62	1.85	2.22
P_{75}	1.68	2.08	2.69	2.65	2.29	2.41	2.86
P_{90}	1.94	2.52	3.39	3.46	2.92	2.83	3.51
P_{95}	2.10	2.81	3.95	4.12	3.35	3.08	3.93
P_{99}	2.45	3.30	4.37	4.62	4.16	3.58	4.73
Grade 7							
Mean	1.62	1.96	2.58	2.62	1.99	2.24	2.83
SD	0.42	0.70	1.01	1.00	1.03	0.70	0.81
Skew.	-0.47	-0.45	0.22	0.45	0.17	0.09	0.54
Kurt.	1.13	1.24	-0.12	0.11	0.31	-0.44	0.45
P_{01}	0.25	0.10	0.53	0.73	-0.29	0.80	1.29
P_{05}	0.93	0.84	1.04	1.17	0.34	1.12	1.65
P_{10}	1.10	1.10	1.25	1.36	0.69	1.31	1.84
P_{25}	1.38	1.54	1.85	1.88	1.30	1.72	2.25
P_{50}	1.64	1.97	2.50	2.50	1.96	2.23	2.76
P_{75}	1.90	2.44	3.25	3.26	2.65	2.75	3.34
P_{90}	2.12	2.83	3.82	3.88	3.31	3.15	3.90
P_{95}	2.25	3.08	4.19	4.31	3.70	3.38	4.25
P_{99}	2.53	3.31	5.14	5.36	4.59	3.84	5.07
Grade 8							
Mean	1.81	2.25	3.06	3.11	2.30	2.62	3.24
SD	0.39	0.70	0.99	0.99	1.09	0.73	0.73
Skew.	-0.66	-1.14	0.44	0.59	0.04	0.19	0.58
Kurt.	1.28	3.28	0.14	0.44	0.31	-0.45	0.66
P_{01}	0.47	-0.29	1.07	1.21	-0.19	1.18	1.86
P_{05}	1.14	1.11	1.57	1.68	0.56	1.49	2.20
P_{10}	1.33	1.43	1.78	1.88	0.91	1.67	2.36
P_{25}	1.59	1.87	2.34	2.39	1.58	2.06	2.72
P_{50}	1.84	2.31	3.02	3.03	2.30	2.60	3.19
P_{75}	2.06	2.72	3.66	3.67	3.00	3.13	3.68
P_{90}	2.27	3.10	4.37	4.40	3.67	3.61	4.19
P_{95}	2.38	3.30	4.89	4.96	4.10	3.86	4.55
P_{99}	2.63	3.30	5.83	6.00	4.97	4.34	5.30

Table 7. Grade-to-Grade Change in Estimated Theta Values for Several Location Indices

Model/Statistic		Grade Level					
		3 to 4	4 to 5	5 to 6	6 to 7	7 to 8	3 - 8
3PL S	Mean	0.66	0.46	0.26	0.24	0.19	1.81
	P_5	1.16	0.76	0.46	0.34	0.21	2.93
	P_{10}	0.93	0.70	0.40	0.29	0.23	2.55
	P_{25}	0.76	0.57	0.33	0.27	0.21	2.14
	P_{50}	0.62	0.44	0.25	0.24	0.20	1.75
	P_{75}	0.53	0.32	0.19	**0.22**	0.16	1.42
	P_{90}	0.45	0.22	0.13	**0.18**	0.15	1.14
	P_{95}	0.35	0.21	0.08	**0.15**	0.14	0.93
3PL G	Mean	0.67	0.55	0.39	0.35	0.28	2.25
	P_5	0.96	0.74	0.43	0.38	0.28	2.78
	P_{10}	0.81	0.67	0.42	0.38	0.33	2.61
	P_{25}	0.71	0.60	0.40	0.39	0.34	2.44
	P_{50}	0.63	0.54	0.37	0.37	0.34	2.24
	P_{75}	0.58	0.48	0.37	0.36	0.28	2.08
	P_{90}	0.55	0.43	0.37	0.31	0.26	1.93
	P_{95}	0.49	0.43	0.38	0.27	0.22	1.79
RM	Mean	0.73	0.67	0.60	0.58	0.48	3.06
	P_5	0.76	0.58	0.48	**0.69**	0.54	3.03
	P_{10}	0.81	0.70	0.52	**0.59**	0.49	3.11
	P_{25}	0.87	0.55	0.46	**0.69**	0.53	3.09
	P_{50}	0.79	0.64	0.61	0.55	0.51	3.10
	P_{75}	0.67	**0.73**	0.64	0.56	0.41	3.01
	P_{90}	0.67	**0.77**	0.70	0.42	**0.56**	3.11
	P_{95}	0.66	**0.84**	0.79	0.24	**0.70**	3.23
PCM	Mean	0.73	0.70	0.61	0.59	0.48	3.11
	P_5	0.78	0.64	0.54	**0.63**	0.51	3.10
	P_{10}	0.87	0.61	0.50	**0.64**	0.52	3.15
	P_{25}	0.80	0.74	0.52	**0.59**	0.51	3.15
	P_{50}	0.78	0.65	0.58	**0.59**	0.54	3.14
	P_{75}	0.68	**0.73**	0.63	0.61	0.41	3.06
	P_{90}	0.67	**0.79**	0.75	0.42	**0.52**	3.15
	P_{95}	0.65	**0.88**	**0.89**	0.18	**0.66**	3.26
GPCM	Mean	0.87	0.84	0.60	0.52	0.41	3.24
	P_5	0.94	0.80	0.73	0.64	0.55	3.66
	P_{10}	0.91	0.82	0.69	0.61	0.52	3.55
	P_{25}	0.93	0.84	0.62	0.58	0.47	3.44
	P_{50}	0.89	0.82	0.58	0.54	0.43	3.26
	P_{75}	0.78	**0.87**	0.56	0.40	0.29	2.90
	P_{90}	0.84	0.82	0.58	0.49	0.33	3.06
	P_{95}	0.71	**0.98**	0.51	0.31	0.30	2.82
GRM	Mean	0.77	0.64	0.44	0.40	0.38	2.62
	P_5	0.80	0.78	0.54	0.52	0.37	3.02
	P_{10}	0.80	0.77	0.53	0.49	0.36	2.94
	P_{25}	0.83	0.74	0.48	0.44	0.34	2.83
	P_{50}	0.80	0.64	0.44	0.38	0.37	2.64
	P_{75}	0.71	0.55	0.42	0.34	**0.39**	2.41
	P_{90}	0.68	0.49	0.36	0.32	**0.45**	2.29
	P_{95}	0.64	0.48	0.32	0.29	**0.48**	2.22
NM	Mean	0.69	0.56	0.42	0.32	0.31	2.30
	P_5	0.77	0.62	0.31	0.23	0.22	2.15
	P_{10}	0.73	0.64	0.30	0.25	0.22	2.15
	P_{25}	0.72	0.62	0.34	0.29	0.29	2.25
	P_{50}	0.69	0.54	0.40	0.34	0.34	2.29
	P_{75}	0.66	0.49	**0.50**	0.36	0.35	2.37
	P_{90}	0.65	0.48	**0.57**	0.39	0.36	2.44
	P_{95}	0.60	0.52	**0.61**	0.35	**0.40**	2.49

Notes. Change values greater than the change value from the previous grade are in bold type.

Table 8. Indices of Variability for Theta Estimates across Grades

Model	Grade Level 3	4	5	6	7	8	8/3 Ratio
	Standard Deviation						
3PL S	1.00	0.78	0.59	0.47	0.42	0.39	0.39
3PL G	1.00	0.86	0.75	0.73	0.70	0.70	0.70
RM	1.00	0.95	**0.99**	**1.08**	1.01	0.99	0.99
PCM	1.00	0.94	**0.97**	**1.07**	1.00	0.99	0.99
GPCM	1.00	0.93	**0.96**	0.89	0.81	0.73	0.73
GRM	1.00	0.94	0.83	0.76	0.70	**0.73**	0.73
NM	1.00	0.95	0.91	**1.00**	**1.03**	**1.09**	1.09
	$(P_{75} - P_{25}) / 2$						
3PL S	0.59	0.48	0.35	0.28	0.26	0.24	0.40
3PL G	0.60	0.54	0.48	0.46	0.45	0.42	0.70
RM	0.71	0.64	**0.65**	**0.71**	0.70	0.66	0.93
PCM	0.68	0.62	0.62	**0.68**	**0.69**	0.64	0.94
GPCM	0.67	0.62	0.61	0.59	0.55	0.48	0.72
GRM	0.74	0.69	0.59	0.56	0.51	**0.53**	0.72
NM	0.65	0.62	0.56	**0.64**	**0.68**	**0.71**	1.09
	$(P_{95} - P_{5}) / 2$						
3PL S	1.62	1.22	0.94	0.75	0.66	0.62	0.38
3PL G	1.59	1.36	1.20	1.18	1.12	1.09	0.69
RM	1.56	1.52	**1.64**	**1.80**	1.58	**1.66**	1.06
PCM	1.56	1.50	**1.62**	**1.79**	1.57	**1.64**	1.05
GPCM	1.59	1.48	**1.57**	1.46	1.30	1.17	0.74
GRM	1.59	1.50	1.35	1.24	1.13	**1.18**	0.75
NM	1.60	1.52	1.47	**1.62**	**1.68**	**1.77**	1.10

Notes. Variability indices greater than the value from the previous grade are in bold type.

Table 9. BILOG-MG Phase 2 Means and SDs for the 3PL Concurrent Group Calibration across Grades

	Grade Level					
	3	4	5	6	7	8
Mean	0.00	0.79	1.45	1.92	2.37	2.76
SD	1.00	0.89	0.79	0.78	0.76	0.72

Figure 1. Boxplots for Q_3 Between and Within Testlets at each Grade

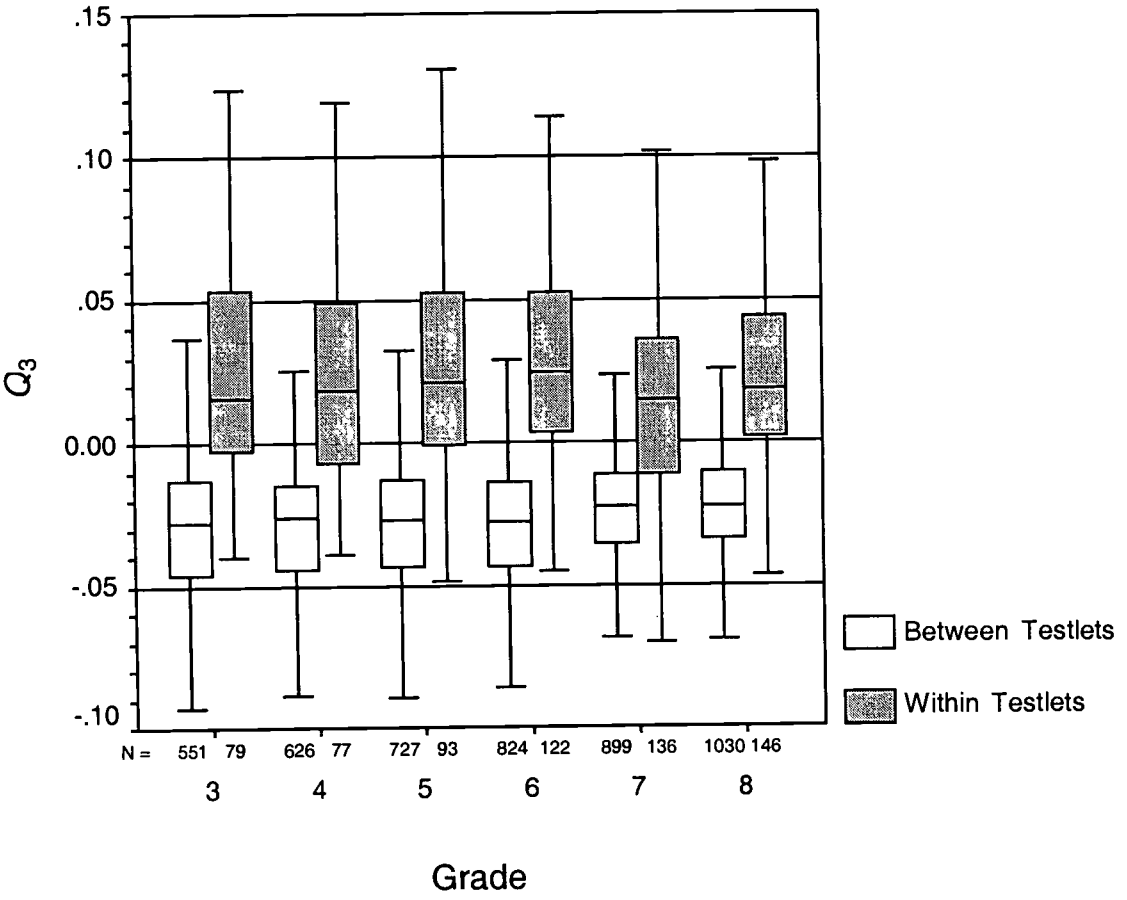
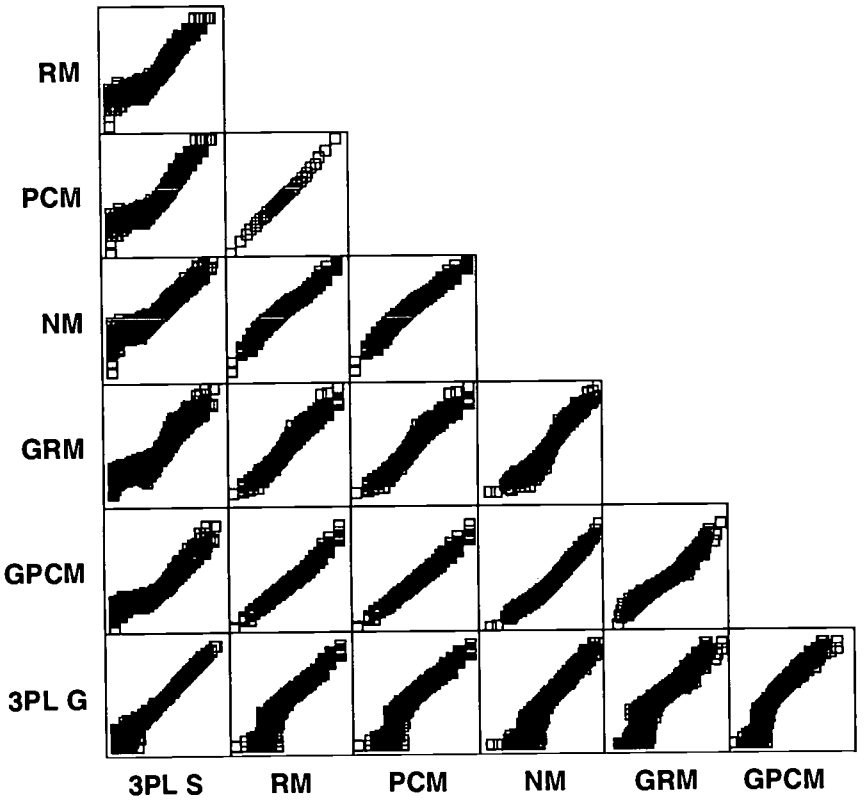


Figure 2. Scatterplots for Estimated Thetas

GRADE 3



GRADE 3 - 8

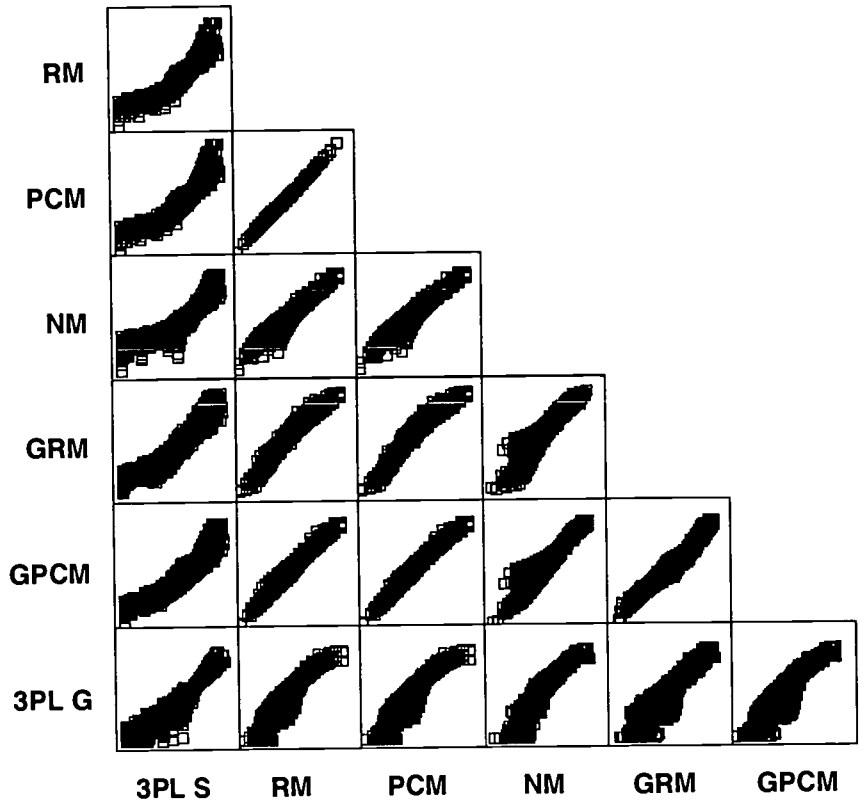


Figure 3. Grade-to-Grade Change in Estimated Thetas at Five Percentiles

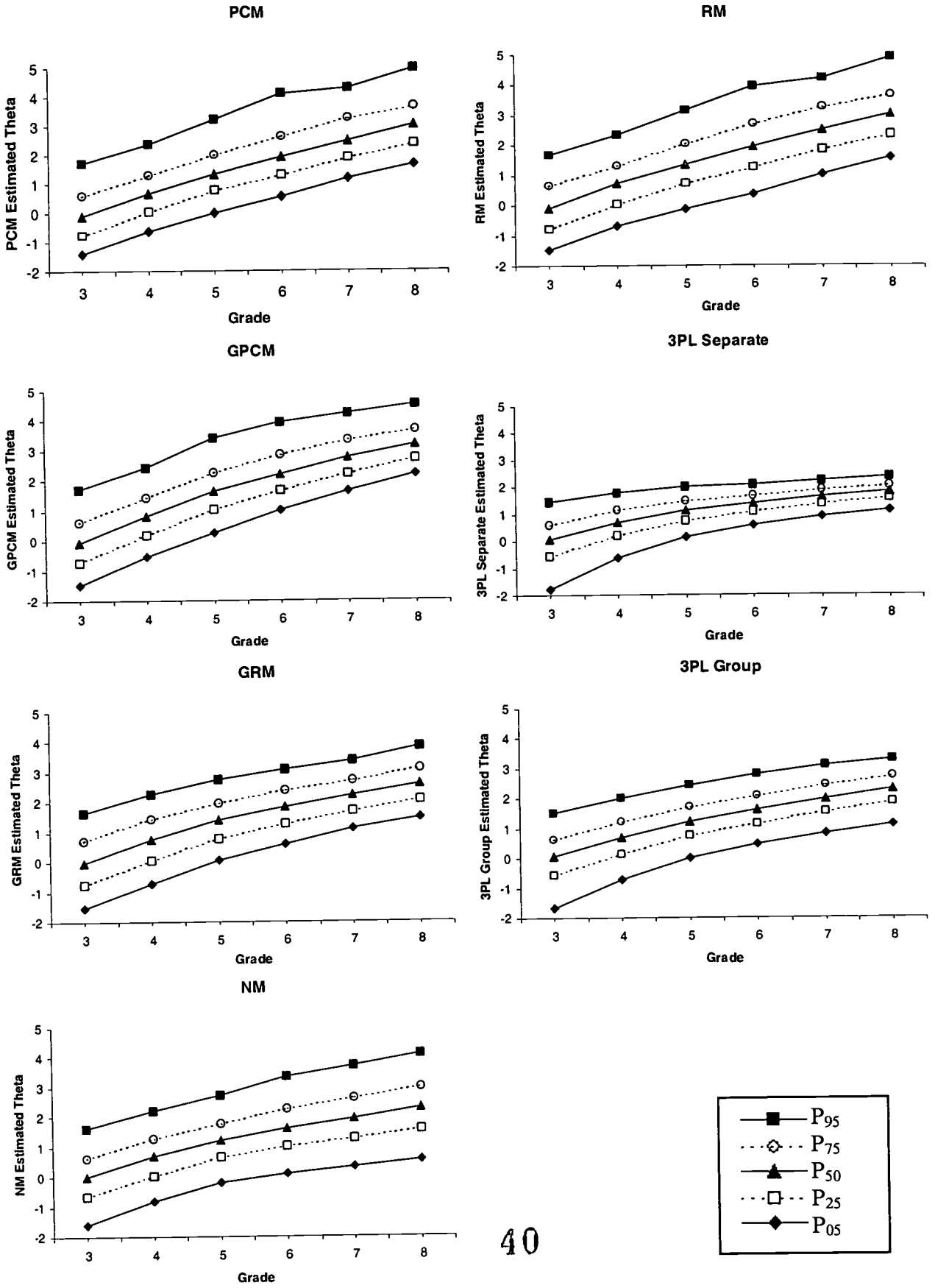


Figure 4. Model Differences in Estimated Thetas for Four Distributional Indices

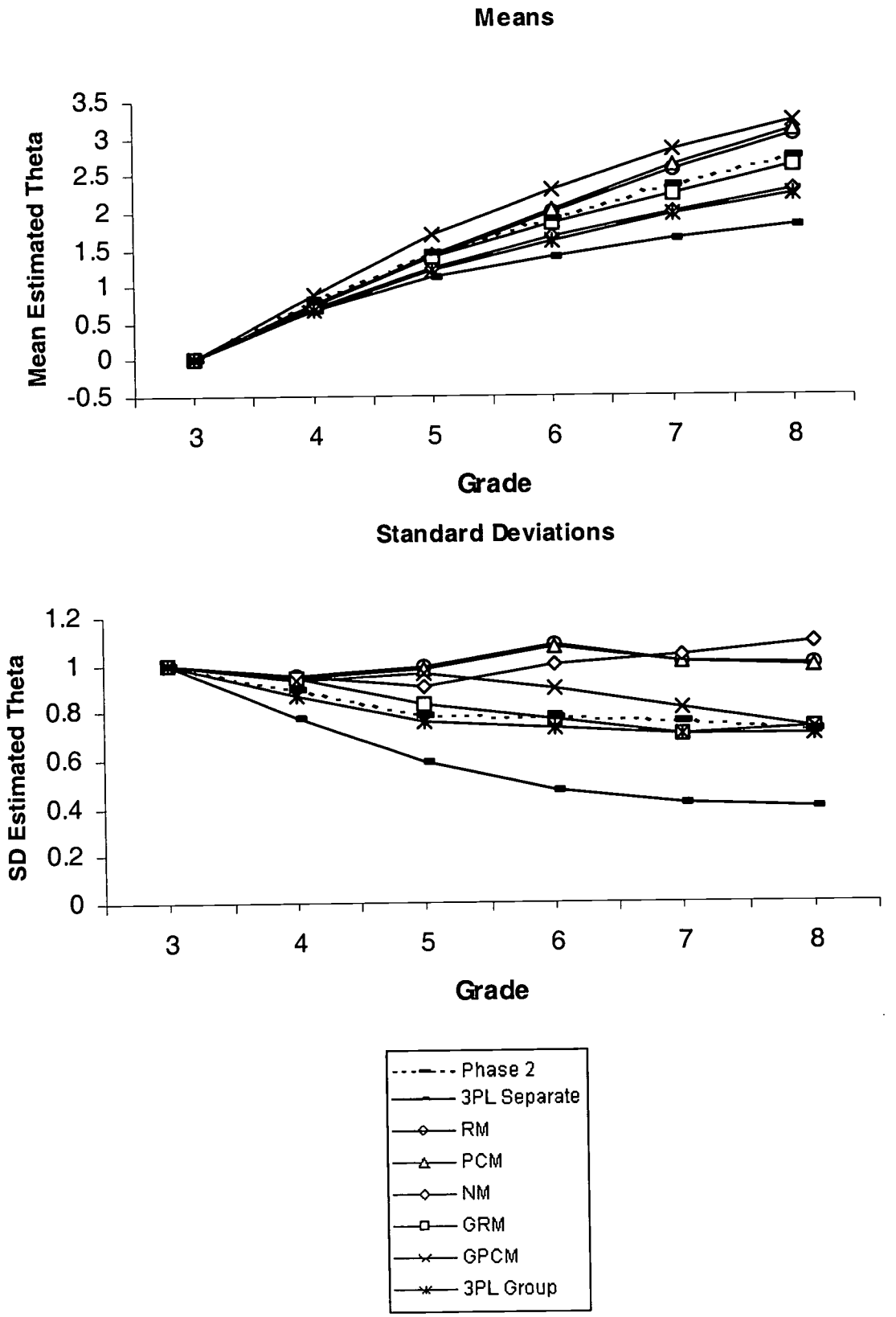
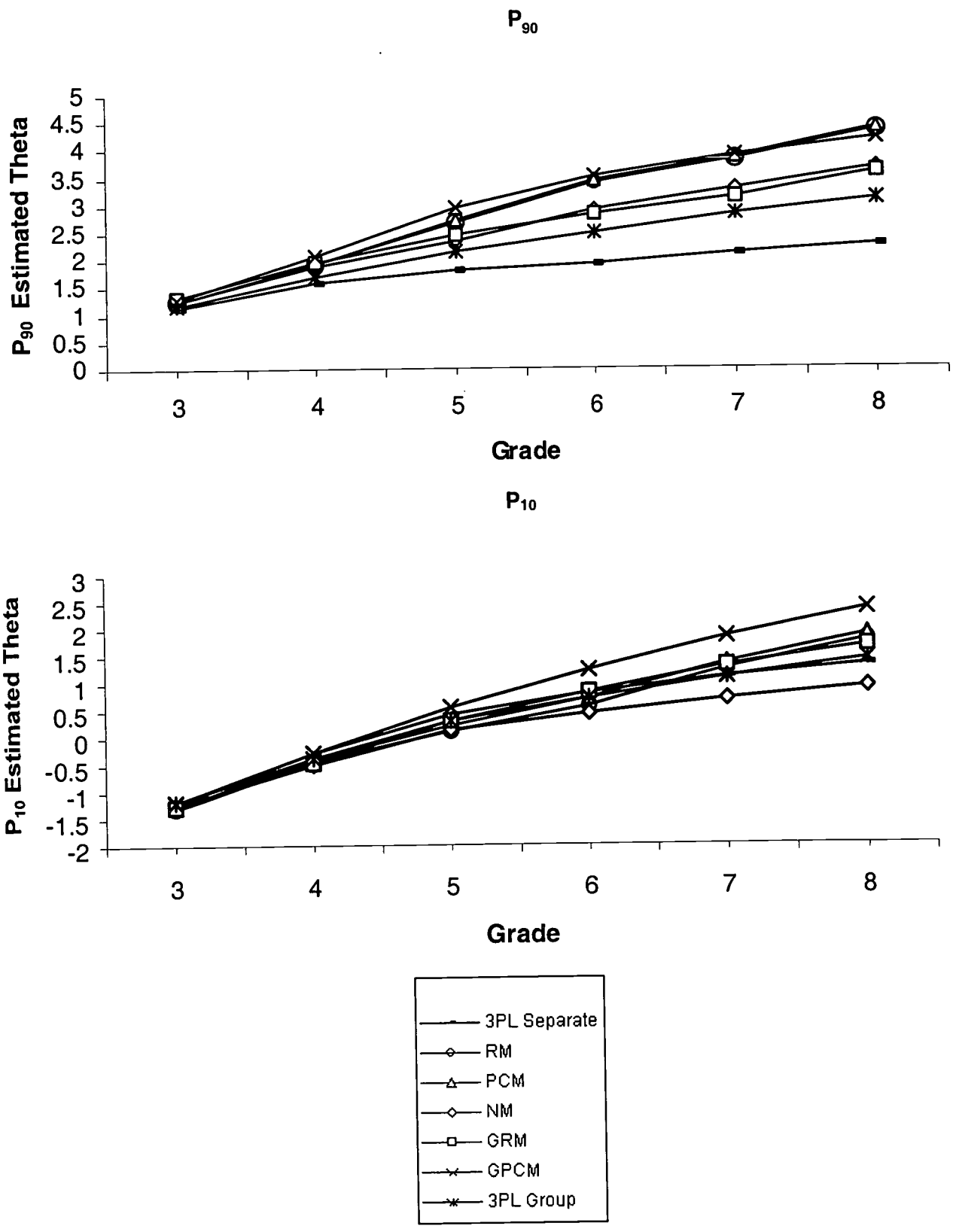


Figure 4. Continued



Author Note

The authors would like to thank Mr. Corey Jackson for his help in providing preliminary descriptive statistics for the data.

TM034249

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: COMPARING VERTICAL SCALES DERIVED FROM DICHOTOMOUS AND POLYTOMOUS IRT MODELS FOR A TEST COMPOSED OF TESTLETS

Author(s): N. SCOTT BISHOP MD HAFIDZ OMAR

Corporate Source:

Publication Date: April 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: N. Scott Bishop	Printed Name/Position/Title: Dr. N. Scott Bishop
Organization/Address: Riverside Publishing, 425 Spring Lake Dr, Itasca IL 60143	Telephone: 630 467 6163 FAX: 630 467 7126
	E-Mail Address: Scott_Bishop@hmco.com Date: 5/31/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)